IN THE UNITED STATES DISTRICT COURT
FOR THE WESTERN DISTRICT OF WISCONSIN

WISCONSIN ALUMNI RESEARCH FOUNDATION, Plaintiff, v. INTEL CORPORATION, Defendant.	Civil Action No. 3:08-cv-00078-bbc DEMAND FOR JURY TRIAL

DEFENDANT INTEL CORPORATION'S
REPLY MARKMAN BRIEF

TABLE OF CONTENTS

TABLE OF AUTHORITIES

Federal Cases

Defendant Intel Corporation ("Intel") submits this reply brief in further support of its proposed claim constructions for the patent-in-suit, U.S. Patent No. 5,781,752 ("the '752 patent"), and in response to the opening Markman brief ("Opening Brief") of plaintiff Wisconsin Alumni Research Foundation ("WARF").

I. BACKGROUND TECHNOLOGY

The parties' discussions of the background technology are notably different. Most importantly, Intel relies on the intrinsic evidence, whereas WARF relies on extrinsic evidence. Since WARF's starting point is misplaced, much of the analysis that follows is flawed.

In the "Background Technology" section of its Opening Brief, Intel discussed the technology in the context of the '752 patent specification and the prior art cited to the Patent Office during prosecution of the '752 patent. The figures and citations in Intel's analysis all come directly from those intrinsic sources. (Intel Opening Br. at 1-19.)

In contrast, WARF's background sections contain no figures from the '752 patent, but rather a series of figures that WARF has created for this litigation. (WARF Opening Br. at 5-20.) Similarly, WARF's background sections discuss concepts that are not discussed anywhere in the '752 patent specification, such as tracking execution by clock cycles (p. 13); "memory hierarchy" (p. 17); and memory order buffers (p. 9). *See* Clark Decl., ¶ 5.[1] The words "clock," "memory hierarchy," and "buffers" do not appear in the '752 patent. *Id.* The named inventors did not address this level of implementation-detail in their patent. *Id.*

By addressing its background sections to issues that were of no concern to the named inventors and that are not mentioned in the patent specification, WARF establishes a false foundation for the arguments that follow. "[E]xtrinsic evidence may be useful to the court, but it

[1] "Clark Decl." refers to the Declaration of Douglas Clark, filed herewith.

is unlikely to result in a reliable interpretation of patent claim scope unless considered in the context of the intrinsic evidence." *Phillips v. AWH Corp.* 415 F.3d 1303, 1319 (Fed. Cir. 2005). Intel's background technolgy section is rooted in the intrinsic evidence; WARF's background technology section overlooks the intrinsic evidence.

II. DISPUTED TERMS

A. The court should construe all nine disputed terms

Claim 1 of the '752 patent contains 3 basic parts:

> (A) a **data speculation circuit** detects a **mis-speculation** by instructions that are **in fact executed**.
>
> (B) a **predictor** creates a **prediction** based on the history of mis-speculations
>
> (C) a **threshold detector** prevents speculation once the mis-speculation rate moves within a **predetermined range**.

See Intel Opening Br. at 15 (describing the "basic sequence").

None of the items unique to part (C) is disputed. This leaves five key terms in dispute: data speculation circuit; mis-speculation; in fact executed; predictor; prediction. Intel proposes to resolve all five; WARF proposes to resolve only two: "in fact executed" and "prediction."

"Data speculation circuit" should be construed for at least two reasons. First, WARF agrees that the Court should resolve the issue of whether the '752 patent ties mis-speculation and prediction to load/store pairs. This issue first arises with respect to "data speculation circuit." Intel submits that it does not make sense to resolve the pair issue with respect to "prediction" (as WARF agrees is crucial), yet not with respect to "data speculation circuit." Second, WARF admits that "in fact executed" should be construed. But "in fact executed" appears in the same clause as "data speculation circuit": the data speculation circuit detects mis-speculation by

instructions that are "in fact executed." Since both terms are linked in the claim, both should be construed.

"Mis-speculation" should be construed for this same reason: as claim 1 is written, mis-speculation is a portion of the same limitation that includes "data speculation circuit" and "in fact executed." All three need to be construed to give the limitation its full meaning.

Similarly, "predictor" should be construed because it is linked to "prediction" (which WARF agrees should be construed). In the claim, the predictor produces a prediction. It does not make sense to construe one without the other.

For reasons stated in its Opening Brief, Intel also believes that four additional terms should be construed: "where a data consuming instruction . . . is in fact executed before the data producing instruction"; "mis-speculation indication"; "prediction associated with the particular data consuming instruction . . ."; and, "a prediction threshold detector. . . ." Intel believes that these four terms should be construed to provide the jury a meaningful charge. (*See* Intel's Motion Requesting That The Court Construe Patent Terms, at 3-4.) Because WARF did not address any of those four extra terms in its Opening Brief, Intel will rely upon its original briefing, and does not address those terms in this reply.

B. The five key terms in dispute

1. "in fact executed"

'752 Claim Language	WARF's Proposed Construction	Intel's Proposed Construction
1. In a processor capable of executing program instructions in an execution order differing from their program order, the processor further having a data speculation circuit for detecting data dependence between instructions and detecting a mis-speculation where a data consuming instruction dependent for its data on a data producing instruction of earlier	**in fact executed:** a LOAD instruction is "in fact executed" before the STORE instruction when the LOAD instruction has actually accessed *or was certain to access data* that has not yet been updated by the STORE	**in fact executed**: a load instruction is "in fact executed" when the load instruction actually has loaded data from a memory location

program order, is **in fact executed** before the data producing instruction, a data speculation decision circuit comprising:	instruction.	

Intel's proposed construction follows the plain meaning of "executed" as further explained in the patent specification. The plain meaning of "in fact executed" is "actually carried out," and the specification is clear that a load instruction is "executed" when it has loaded the value from the memory location. (9:8-15; *see also* 7:18-20, 2:36-40; *see generally* Intel Opening Br. at 23-27.) WARF complains that Intel's proposal "[does not] capture the full scope of this term." (WARF Br. at 32.) In fact, WARF is trying to incorrectly expand the scope of the claim by adding an alternative to the claim language (with the phrase "or was certain to access data"), one that is contrary to the plain meaning and to the patent specification.

WARF's brief and supporting expert declaration of Professor Dally confirm that WARF may ultimately be seeking to change "in fact executed" to mean "in the process of executing" or "partly executed." *See, e.g.*, WARF Br. at 33 ("bound to access"); at 33 ("started the process"); at 34 ("being performed"); at 36 ("being performed"); Dally Decl. ¶ 67 ("at least partly executed"); Dally Decl. ¶ 68 ("starts execution"). WARF's proposed construction is thus contrary to the patent specification and to the words used in the patent claim. *See* Clark Decl., ¶7.

Claim 1 says "in fact executed," and mentions nothing about "in the process of executing" or "partly" executed. *See* Clark Decl., ¶8. Nor does the claim or specification use the term that WARF actually proposes, "was certain to access data." Nor does anything in the patent specification even indicate that an instruction has "in fact executed" once it is "certain" to access data. *Id.* Indeed, the only citations to the patent specification that WARF proffers (7:1-7; 9:61-64) use the terms "performed" and "executed" in the past tense. (WARF Opening Br. at 36.)

See also Clark Decl., ¶ 8. WARF's proposed construction is thus contrary to the most basic rules of claim construction.

> The best source for understanding a technical term is the specification from which it arose. . . . [I]n case of doubt or ambiguity it is proper in all cases to refer back to the descriptive portions of the specification to aid in solving the doubt or in ascertaining the true intent and meaning of the language employed in the claims . . . . The construction that stays true to the claim language and most naturally aligns with the patent's description of the invention will be, in the end, the correct construction.

See Phillips, 415 F.3d at 1315-16 (internal citations, quotations, and edits omitted).

In its Opening Brief, WARF created two figures—nowhere in the patent—that purport to show what could happen if a dependent load instruction "takes a very long time" to execute. (WARF Br. at 37-38, and Figs. 11, 12.) *See* Clark Decl., ¶ 9. In Figure 11, WARF posits an example of a load instruction that starts to execute at clock cycle 5, but does not finish execution ("in fact executed") until cycle 97. *See* Clark Decl., ¶ 9. In the meantime, the store instruction has executed at cycle 13. *See* Clark Decl., ¶ 9. WARF then concludes, "such a system would not detect the mis-speculation, and would not squash this incorrectly executed load instruction." (WARF Br. at 37.) WARF shows an allegedly better system in Figure 12. (*Id.* at 38.) *See* Clark Decl., ¶ 9. There are at least two problems with WARF's examples. *See* Clark Decl., ¶ 9.

First, the examples and figures are an invention of WARF for the purposes of litigation, not examples or figures from the '752 patent. *See* Clark Decl., ¶ 10. Notably, this section of WARF's brief is devoid of a single citation to the patent specification or file history to support its construction. WARF's argument has disconnected itself from the intrinsic evidence, which is an error.

Second, WARF's examples posit a "straw man" problem that would not actually exist under Intel's proposed construction. *See* Clark Decl., ¶ 10. As described in the accompanying

declaration of Dr. Douglas Clark, WARF's Figure 11 shows a "problem" that no person of ordinary skill in the art would allow to happen in the first instance. *See* Clark Decl. ¶ 10.

Regardless, even presuming that the scenario shown in WARF's Figure 11 could cause a mis-speculation under Intel's proposed construction of "in fact executed," the resulting mis-speculation would be squashed. *See* Clark Decl., ¶ 11. Element 76 of Figure 3 of the '752 patent shows that the processor analyzes load instructions that have already executed, and then either squashes them (if a previously executed store was to the same address) or retires them (if no previously executed store instruction was to the same address). (Fig. 3 and 10:26-34.) *See* Clark Decl., ¶ 10. WARF alleges concern that the processor would not recognize and squash an error (WARF Br. at 37), but this argument overlooks element 76 in Figure 3. *See also* 7:31-65 (explaining that after execution is completed, the instruction is either squashed or retired).

Additionally, the level of detail posited in WARF's Figures 11 and 12 is absent from the claim language and from the patent specification. *See* Clark Decl., ¶ 12. The implementation-level details of how instructions execute and how they are then either squashed or retired are not part of the claim language and are not discussed in meaningful detail in the specification. *Id.*

WARF's Figure 12 (WARF Br. at 38) is similarly disconnected from the intrinsic evidence. *See* Clark Decl., ¶ 13. WARF's Figure 12 shows how a load instruction that is in the process of executing could be squashed "right away" (WARF Br. at 38) rather than waiting for it to be "in fact executed" and then squashed later. *See* Clark Decl., ¶ 13. According to WARF, squashing the instruction right away "makes sense." (WARF Br. at 38.) Maybe so. But the claim language states that the load instruction at issue is one that has "in fact executed," and the specification says nothing about WARF's scenario. *See* Clark Decl., ¶ 13. It might "make sense" to identify imminent mis-speculations that are about to be caused by instructions that have

only partly executed, but that is not a function that the specification describes, or that claim 1 recites. Inventions are defined by their claims, not by what a litigant decides more than a decade later would "make sense."

In sum, WARF's Figures 11 and 12 are not related to the intrinsic evidence, posit an incorrect solution to a non-existent problem, and should be disregarded. *See Phillips*, 415 F.3d at 1318 ("a court should discount any expert testimony that is clearly at odds with the claim construction mandated by the claims themselves, the written description, and the prosecution history") (internal quotation omitted).

2. "data speculation circuit"

'752 Claim Language	WARF's Proposed Construction	Intel's Proposed Construction
1. In a processor capable of executing program instructions in an execution order differing from their program order, the processor further having a **data speculation circuit** for detecting data dependence between instructions and detecting a mis-speculation where a data consuming instruction dependent for its data on a data producing instruction of earlier program order, is in fact executed before the data producing instruction, a data speculation decision circuit comprising:	**data speculation circuit:** a circuit that detects data *dependence between LOAD and STORE instructions* and tracks execution of such instructions in order to detect any mis-speculations arising from the data speculative execution of LOAD instructions.	**data speculation circuit:** a circuit that detects data *dependence between load/store pairs* and detects a mis-speculation by a load instruction that is in fact executed

The parties dispute whether the mis-speculation that is identified and which then forms the basis of the prediction is (or is not) based on a load/store **pair**. In its Opening Brief, Intel briefed this issue in the context of the term "data speculation circuit," whereas WARF briefed the issue in the context of the term "prediction." In either case, the issue is the essentially the same.

a. The plain meaning of the claim supports inclusion of "pairs"

As explained in Intel's Opening Brief, while the word "pair" does not appear in claim 1, the plain reading of the claim shows that the instructions at issue are a store instruction and a dependent load instruction, i.e., what the specification calls a "pair."

WARF argues that the language of claim 1 "explicitly refers to a prediction associated with a LOAD instruction, not with a LOAD/STORE pair." (WARF Br. at 24.) Citing claim element 1(a)—which recites a "prediction associated with the particular [load]—WARF states, "here they just refer to a LOAD." (WARF Br. at 25.)

WARF's argument reads only part of the claim in isolation, while ignoring the other parts of the claim that prove Intel to be correct. *See* Clark Decl., ¶ 15. The "particular" load referenced in element 1(a) is the same load instruction that the preamble has **paired** with "the" particular store instruction on which the load depends. (14:40-43.) *See* Clark Decl., ¶ 15. Thus, while the particular part of the claim quoted by WARF does refer to a "particular" load, the remainder of the claim clarifies that the load instruction is "particular" because it is paired with a particular store instruction.

At page 25 of its Opening Brief, WARF makes a quasi-claim differentiation argument by comparing claim 1 to claims 5 and 9. WARF contrasts claims 5 and 9 to claim 1 on the ground that claim 5 and 9 state that the store is "associate[ed] with" the load, whereas claim 1 does not. (WARF Opening Br. at 25.) Again, WARF's argument overlooks the whole text of claim 1, which unequivocally "associates" a load and store to each other ("pair") in the preamble by saying that the load instruction is dependent on the store instruction. The fact that claim 1 does not use the word "associating" or "associated" is irrelevant. *See Curtiss-Wright Flow Control Corp. v. Velan, Inc.*, 438 F.3d 1374, 1381 (Fed. Cir. 2006) ("a patentee may define the same subject matter with claims having different terminology").

b. The specification says that "pairs" are the key to the "invention"

As explained in Intel's Opening Brief, the specification repeatedly says that "pairs" are part of "the invention." (Intel Opening Br. at 29-30.) Claims are construed to comport with "the invention" as defined in the specification. *See, e.g.*, *Phillips*, 415 F.3d at 1316 (claims are construed to cover "what the inventors actually invented"). Hence, "load/store pairs" is a necessary limitation on the claims.

WARF ignores the intrinsic evidence that refers to "the invention" as including load/store pairs. *See* Clark Decl., ¶ 16. WARF's expert Professor Dally, for example, notes that the patent specification "describes" a "prediction table" and "predictions based on load/store pairs of the preferred embodiment." (Dally Decl. ¶ 60.) But this observation accounts for only half of the evidence. *See* Clark Decl., ¶ 16. The patent specification does not merely "describe" load/store pairs; rather, it describes them as being part of "the invention." *See* Clark Decl., ¶ 16. Similarly, the load/store pairs are not merely the "preferred embodiment" as Professor Dally says, but rather are the **only** embodiment. *See* Clark Decl., ¶ 16. Under this fact pattern, the Federal Circuit consistently holds that the claim terms should be limited to what the specification describes as the invention. *See, e.g., Honeywell Int'l, Inc. v. ITT Indus., Inc.*, 452 F.3d 1312, 1318-19 (Fed. Cir. 2006) (limiting claim term to the feature that the specification referred to "[o]n at least four occasions" as constituting the "invention"); *Watts v. XL Systems, Inc.*, 232 F.3d 877, 882-83 (Fed. Cir. 2000) (limiting claim term to the feature that was described as part of "[t]he present invention" and that was the only disclosed embodiment); *Honeywell Int'l, Inc. v. Universal Avionics Sys. Corp.*, 493 F.3d 1358, 1362 (Fed. Cir. 2007) (limiting claim term to a particular feature, where the specification "discloses no other" embodiment and also described

the feature as an "important feature of the present invention"). *See also* Intel Opening Br. at 30, 32 (citing additional cases).

Notably, in at least two places in its Opening Brief, WARF (perhaps unavoidably) acknowledges that pairs **are** part of the alleged invention. *See* Clark Decl., ¶ 17.

First, WARF asserts that the prior art '506 patent is inferior to the '752 patent because the '506 patent "unnecessarily blocks LOAD instructions that do **not** conflict with **the offending** STORE," and thus lose performance by blocking speculative execution for too many load instructions. (WARF Br. at 18, citing Dally Decl. ¶ 42, emphasis added.) In other words, according to WARF, the '752 patent is an advance over the prior art because the '752 patent blocks speculative execution of load instructions that "conflict" with "offending" store instructions, i.e., the store that forms the other half of the load/store pair. *See* Clark Decl., ¶ 17. Thus, in distinguishing prior art, WARF argues that the '752 is different because it prevents data speculation based on load/store pairs. *Id.*

Second, WARF extols the '752 patent by arguing that it "intelligently decide[s]" to perform data speculation because it is "rooted in a phenomenon that . . . was recognized by the inventors," namely that "if a LOAD instruction conflicts with a STORE instruction, then that LOAD instruction is highly likely to conflict **with the same STORE instruction again**." (WARF Br. at 20, emphasis added.) In other words, according to WARF, the '752 patent is "intelligent[]" because it identifies load/store pairs that mis-speculate, and recognizes that if a pair mis-speculates once, it is likely to do so again. *See* Clark Decl., ¶ 18. WARF's statement is an accurate summary of the patent specification, and it proves Intel's point: "the invention" consists of identifying load/store pairs that mis-speculate, and disabling them from speculating. *See* Clark Decl., ¶ 18.

c. The specification shows that the named inventors did not contemplate the "tier" structure now advanced by WARF

The specification shows that the named inventors did not contemplate the alleged invention discussed by WARF in its Opening Brief. In particular, the named inventors did not contemplate the "tier" structure that WARF discusses in its brief.

As explained in Intel's Opening Brief, there is nothing in the intrinsic evidence that indicates that the inventors contemplated or intended to claim a processor that controlled data speculation other than with load/store pairs. *See* Clark Decl., ¶ 19. Every figure and example in the '752 patent concerns identifying and handling a mis-speculating load instruction that is paired with a particular store instruction. (Intel Opening Br. at 30-32.) *See* Clark Decl., ¶ 19.

WARF repeatedly refers to this structure as the "preferred embodiment." (E.g., WARF Br. at 26, "one embodiment of the invention"; at 29, "preferred embodiment"; at 30, "one embodiment.") However, this is the **only** embodiment. *See* Clark Decl., ¶¶ 16, 19. Thus, the specification as a whole confirms that the alleged invention is limited to a processor that controls data speculation based on load/store pairs. *Id.*

WARF's argument to the contrary incorrectly reads the '752 patent specification. According to WARF, the "invention" is divided into three tiers: in tier one, instructions can speculate; in tier two, particular load instructions are prevented from mis-speculating based on their prediction value (but without reference to the store instruction); in tier three, a synchronization table is used to determine when the delayed load instruction can execute. According to WARF, claim 1 of the patent is limited to tier one and tier two, and—conveniently for WARF—only tier three is limited to load/store pairs. (WARF Br. at 28-29.) WARF's argument is incorrect.

(i) The patent never defines the three tiers

The "three-tiered approach" is mentioned once in the patent, in passing (3:64). Contrary to WARF's Opening Brief, the patent does not explain or clarify what it means by the "three-tiered approach." *See* Clark Decl., ¶ 20. There is nothing in the patent specification that ties portions of the disclosed system, or the claims, to a certain "tier" or another. *See* Clark Decl., ¶ 20. Similarly, nothing in the patent indicates that claim 1 refers only to "tier one" and "tier two," and nothing indicates that tiers one and two do **not** relate to load/store pairs. *See* Clark Decl., ¶ 20.

For example, the fact that the synchronization table is used in the so-called "tier three" (4:5-7) does not exclude its use from the so-called "tier two." *See* Clark Decl., ¶ 21. Indeed, the only structure that is shown to prevent mis-speculation—i.e., to cause a load instruction to "wait"—is the synchronization table. *See* Fig. 4 elements 116 and 102 (both of which follow from analysis performed with the synchronization table at elements 106, 108, 110, 118, and 122). *See* Clark Decl., ¶ 21. On the one hand, WARF equates "delay[ing]" a load instruction in tier two to "prevent[ing]" data speculation (WARF Br. at 28). On the other hand, WARF alleges that use of the synchronization table is reserved only for "tier three" (WARF Br. at 21-22, alleging that tier three adds "additional hardware" that includes the synchronization table). WARF's problem is that the synchronization table is the only disclosed feature that actually causes a load instruction to "wait," i.e., that actually prevents speculation. *See* Clark Decl., ¶ 21. This is not to say that the synchronization table is, or should be, part of any particular "tier." Nor is it to say that only the use of the "synchronization table" would "prevent[] data speculation" within the meaning of claim 1. *See* Clark Decl., ¶ 21. Rather, the point is that WARF's allocation of features and functionality to the various tiers, and its assertion that claim 1 is limited to tier one

and tier two, are creations of WARF for this litigation, and are not supported by the patent specification or claims. *See* Clark Decl., ¶ 21.

In sum, WARF starts with the result it wants, namely to eliminate the "store" half of the pair from claim 1. WARF then takes various claim elements and the various features from the specification, and declares—without support in the intrinsic evidence—to which "tier" those elements/features belong. WARF's argument mis-reads and mis-applies the '752 patent's mention of the "three-tiered approach."

(ii) WARF is incorrect that the patent ties prediction to just the load instruction

WARF argues the '752 patent involves the processor identifying a mis-speculating **load** instruction and using a prediction table to develop a prediction value for that load instruction. *See* Clark Decl., ¶ 22. According to WARF, each time a load is encountered, the prediction table is searched for **that load**, and a determination is made with respect to speculative execution of that load. (WARF Opening Br. at 19-20; Dally Decl. ¶¶ 46-48.)

WARF's analysis disregards the other half of the pair: the particular **store** instruction that corresponds to the particular load instruction at issue. *See* Clark Decl., ¶ 23. WARF's analysis is akin to describing how to walk by referring only to the movements of the left foot: it is only half the story.

In fact, the prediction value does not express the likelihood of a load mis-speculating on its own; instead, the prediction value expresses the likelihood of a load mis-speculating with respect to a particular store, and not with respect to any other store. *See* Clark Decl., ¶ 23. The system is **not** concerned with a load instruction that mis-speculates too often; the system is concerned with a load instruction that mis-speculates too often **when paired with a particular store.** *See* Clark Decl., ¶ 23. It is only mis-speculation by a load/store pair that is logged in the

prediction table and assigned a prediction value. *Id.* Indeed, this concern is the (alleged) central insight of the patent, namely that "most data dependence mis-speculations can be attributed to a few static STORE/LOAD pairs." (3:51-53.) *See also, e.g.*, Fig. 5, element 109 (prediction value is for a pair); 11:8-14 (prediction table contains an entry for a load and its partner store); 11:30-33 (prediction value is for a particular load and a particular store); 11:44-46 (prediction table "compiles a prediction statistic as to whether data dependence exists between a LOAD/STORE pair"); 12:65 to 13:3 (prediction table is checked for "the load/store pair," and updated "so that **this** mis-speculation may be avoided in the future"); 14:1-3 ("the prediction is used to determine the likelihood of a dependency between **two instructions** in the future"). *See* Clark Decl., ¶ 23.

Stated differently, the (alleged) central insight of the '752 patents was to keep track of **pairs**, because (according to the inventors) it was the pairs that were causing "most" (3:51) of the problems. *See* Clark Decl., ¶ 24. By design, this meant **not** keeping track of mis-speculations by loads individually. *See* Clark Decl., ¶ 24. Indeed, general tracking of all data dependencies was deemed "overwhelming" (3:44) by the named inventors, and hence excluded from the '752 patent. *See* Clark Decl., ¶ 24. WARF's current assertions thus contradict what the named inventors said their invention was. *See* Clark Decl., ¶ 24.

It is true that the '752 patent refers to a load instruction that mis-speculates, and to a prediction value for a load instruction. *See, e.g.*, 4:2 ("mis-speculations for that LOAD instruction"); 10:19-22 ("prediction as to whether the LOAD should take place"). But for each of these passages that WARF cites, the patent goes on to explain (in passages that WARF omits) that the load instruction is considered and handled as part of a **pair** of instructions. *See* Clark Decl., ¶ 25. Thus, after referring to a prediction for a LOAD instruction at column 4, line 2, the patent explains what it means in the next paragraph: "Specifically, the present invention provides

a speculation decision circuit . . . . The speculation decision circuit includes a predictor circuit receiving the mis-speculation signal from the data speculation circuit to **produce a prediction associated with the particular data producing/consuming instruction pair** and based on the mis-speculation indication." (4:8-21) (emphasis added). *See* Clark Decl., ¶ 25. Similarly, after referring to a prediction for a load instruction at column 10, lines 19-22, in the context of a "READY TO LOAD request," the patent explains what it means in the sub-section entitled "Handle Ready to Load." (10:58 to 12:19.) *See* Clark Decl., ¶ 25. That sub-section explains the prediction table, which has entries for **pairs** of instructions (11:8-14), with predictions based on the **pair** of instructions (11:30-33; 11:42-45). *See* Clark Decl., ¶ 25. Further, in other places the patent refers to mis-speculation by a **pair** of instructions. *See, e.g.*, 3:51-53 ("mis-speculations can be attributed to a few static LOAD/STORE pairs"); 3:55 ("one load/store pair causes a data mis-speculation"); 12:65-67 ("the prediction table is checked to see whether the load/store pair causing the mis-speculation is in the prediction table already"). *See* Clark Decl., ¶ 25.

Thus, while it is correct, in part, to say that a load instruction mis-speculates, the full context of the patent reveals that, from the perspective of the '752 patent, the load mis-speculates **as part of a load/store pair**. *See* Clark Decl., ¶ 26. WARF takes the intrinsic evidence out of context to suggest that the only issue is the load instruction alone, i.e., only half of the pair. *See* Clark Decl., ¶ 26. WARF's attempt to expand claim 1 to treatment only of a load instruction: (a) finds no support in the patent specification, (b) is contrary to the language of claim 1 (which refers to both a load and the store on which it depends), and (c) perhaps most importantly, contradicts the alleged insight that animated the patent. *See* Clark Decl., ¶ 26.

(iii) The details of WARF's "tier" argument are incorrect

As discussed above: (i) WARF mis-reads the '752 patent to develop a "tier" theory that places certain functionality into certain tiers according to WARF's liking, rather than according to the actual intrinsic evidence; and, (ii) WARF's ultimate goal is to tie prediction in the claim to just the load instruction, to the exclusion of the store instruction and the load/store pair. The error in WARF's arguments is discussed above. However, WARF's error causes it to mis-interpret the intrinsic evidence, and to make additional errors throughout its brief. For the sake of completeness, the following chart corrects the additional mis-statements by WARF about the features and functions of the '752 patent.

WARF argument	**Correction**
"In [tier two] the processor checks the prediction associated with that LOAD instruction [*citing* 3:67 to 4:5]. Based on the value of the prediction, the processor decides either to speculatively execute the instruction or to delay the instruction." (WARF Br. at 21 and Dally Decl. ¶ 46.)	WARF ignores the paragraphs that immediately precede and follow the cited passage, both of which state that the alleged invention is limited to detecting and preventing mis-speculation by load/store pairs. *See, e.g.*, 3:51-53 ("mis-speculations can be attributed to a few static LOAD/STORE pairs"); 4:19-21 ("a prediction associated with the particular [load/store] instruction pair"). *See* Clark Decl., ¶ 28. The processor is "checking" the prediction associated with the load, but that prediction is for a **pair**, not just the load instruction alone. *See* Clark Decl., ¶ 28.
Dally Decl. ¶¶ 47-48 (stating that when a load mis-speculates, it is entered into the prediction table; the prediction value in the table increments each time the load mis-speculates; the load is delayed if the prediction value gets too high)	Professor Dally seeks to tie the prediction value to the load instruction alone, to the exclusion of the store instruction. *See* Clark Decl., ¶ 29. But every passage cited by Professor Dally in ¶¶ 47-48 of his Declaration refers to the treatment of **pairs**, e.g.: 11:29-33 ("The higher the prediction, the more likelihood of mis-speculation if the instruction in the first column [i.e., the load] is executed before the instruction of the second column [i.e., the store]." 12:64-67 ("[T]he prediction table is checked to see whether

WARF argument	Correction
	the LOAD/STORE pair causing the mis-speculation is in the prediction table already.") 13:41-44 ("[T]he addresses of the LOAD and STORE instructions are inserted in the prediction table and the prediction is set to the default value . . . .") *See* Clark Decl., ¶ 29.
During the "third tier" of the invention, the processor "saves the identity of the STORE instruction in the prediction table." (WARF Br. at 21.)	WARF's argument is that the STORE is loaded into the table separately from the load, i.e., not as a pair. *See* Clark Decl., ¶ 30. The specification contains no support for this reading, and unsurprisingly WARF cites no intrinsic evidence in support of its assertion. *See* Clark Decl., ¶ 30.
If the processor only operates tier one and tier two (i.e., not tier three), then the processor "cannot identify the time when the STORE causing the data dependence has executed. In such a processor, the delayed LOAD instruction would be released after all the pending STORE instructions have executed." (WARF Br. at 22.)	The specification contains no support for this reading, and unsurprisingly WARF cites no intrinsic evidence in support of its assertion. *See* Clark Decl., ¶ 31. As noted above, nothing in the patent assigns functions of this particularity to the tiers. *See* Clark Decl., ¶ 31.
"[T]he decision whether to speculatively execute a LOAD instruction or not implicates the LOAD instruction and its associated prediction. It does not involve the STORE instruction of a LOAD/STORE pair." (WARF Br. at 26.)	The specification contains no support for this reading, and unsurprisingly WARF cites no intrinsic evidence in support of its assertion. *See* Clark Decl., ¶ 32. There is no disclosed embodiment in which a load has an "associated prediction" that is separate from the prediction for a load/store pair. *See* Clark Decl., ¶ 32. For example, in decision block 100 in Fig. 4, "the predictor circuit reviews a prediction table show generally in Fig. 5. . . ." (11:4-5.) *See* Clark Decl., ¶ 32. In turn, the prediction table in Fig. 5—which is the only prediction table disclosed in the patent—logs prediction values for load/store **pairs**, not individual load instructions. *See* Clark Decl., ¶ 32.
"[T]he predictor circuit makes a	The passage that WARF cites (10:7-11) concerns the

WARF argument	Correction
prediction as to whether the LOAD should be executed or delayed, and not regarding a LOAD/STORE pair." (WARF Br. at 26, citing 10:7-11, which provides: "If at decision block 48, the instruction received by the data speculation circuit is a LOAD instruction, then at decision block 66 it is determined whether this is a data speculative LOAD, that is whether there are prior STORE instructions on which it might depend.")	decision about whether the execution of the load will or will not be speculative. *See* Clark Decl., ¶ 33. The passage does not discuss or concern prediction or mis-speculation, so the issue of instruction pairs does not even come up. *See* Clark Decl., ¶ 33. Instructions are handled as sets that the patent specification refers to as "windows." *See* Clark Decl., ¶ 33. In Table I, instructions I1 through I3 "represent the instruction window." (7:49.) *See* Clark Decl., ¶ 33. The "window" of instructions is allocated as a group for execution. (10:59-65.) *See* Clark Decl., ¶ 33. The passage that WARF cites (10:7-11) refers to the processor checking to see if there are any unexecuted store instructions of unknown address in the window that precede the load in program order, i.e., "prior STORE instructions on which [the load] might depend." (10:10-11.) *See* Clark Decl., ¶ 33. If not, then there are no unexecuted store instructions on which the load could depend. *See* Clark Decl., ¶ 33. Since there is no possible dependence, there is no speculation, and the load instruction can immediately execute: if the result of decision block 66 of Fig. 3 is "no," then the result is "Handle Load" in block 68. *See* Clark Decl., ¶ 33. The load "has now been released for execution." (13:54). *See* Clark Decl., ¶ 33.
"Figure 4 describes the operation of the predictor circuit. In block 100 of that figure, the predictor checks whether the LOAD instruction is in the prediction table, and not the LOAD/STORE pair." (WARF Br. at 27 and Dally Decl. ¶ 58.)	WARF cites only half of the relevant evidence. *See* Clark Decl., ¶ 34. In element 100, the processor checks the prediction table to see if the load is in the table. *See* Clark Decl., ¶ 34. In the next element—element 104, which WARF overlooks—the processor checks the prediction value of the load/store **pair** to see if the value is high enough to prevent speculative execution **by the pair**. *See* 11:29-33 ("The higher the prediction, the more likelihood of mis-speculation if the instruction of the first column [the load] is executed before the instruction of the second column [the store].") *See* Clark Decl., ¶ 34. The sequence is thus premised on load/store pairs, contrary to WARF's argument. *See* Clark Decl., ¶ 34.

WARF argument	**Correction**
"At process block 76 in Fig. 3, a delayed LOAD instruction waits until it is squashed or until [it] is no longer data speculative [citing 10:25-34]. . . . If all of the previous STORE instructions were for addresses different than the address for the LOAD instruction, then the LOAD instruction is no longer data speculative and can be executed." (WARF Br. at 27.)	WARF overlooks the fact that by the time the processor reaches block 76 of Fig. 3, the load instruction has **already executed**. *See* Clark Decl., ¶ 35. Execution has occurred in element 74. *See* Clark Decl., ¶ 35. As explained in Intel's Opening Brief, instructions that are executed are later either squashed or retired. *See* Clark Decl., ¶ 35. If an executed instruction turns out to be a mis-speculation, the instruction is squashed; if the executed instruction turns out to have speculated correctly, then the instruction is retired. (Intel Opening Br. at 12 (discussing prior art); at 25 (discussing '752 specification).) *See* Clark Decl., ¶ 35. WARF **agrees** with this analysis in its Opening Brief, although it uses the term "commit" in place of "retire." *See, e.g.*, WARF Opening Br. at 13 (instructions "commit" after they execute); at 14-16 ("At a later stage, the processor verifies the correctness of its speculation," and squashes instructions that have mis-speculated). WARF's analysis of block 76 of Fig. 3 is incorrect because it overlooks the fact that the decision to squash/retire comes **after** the instruction has executed. *See* Clark Decl., ¶ 35.
At p. 29 of its Opening Brief, WARF asserts that the purpose of the load/store pair is to "facilitate implementation of the third tier," and is "not required for prediction of mis-speculation." (WARF Opening Br. at 29, citing Dally Decl. ¶¶ 59-60; *see also id.*, "if a processor embodies only tier 1 and tier 2, then the processor could maintain the predictions only for LOAD instructions without implicating the STORE instructions.")	The specification contains no support for this reading, and unsurprisingly WARF cites no intrinsic evidence in support of its assertion. *See* Clark Decl., ¶ 36. Instead, the portions of the specification that WARF cites all support Intel's reading of the patent: the load/store **pair** is used to identify mis-speculations and to update the prediction, without regard to the synchronization table or the so-called "third tier." *See* 11:30-33; 14:3-6; 12:65 to 13:1; 14:15-18. *See* Clark Decl., ¶ 36.

d. WARF's argument about "pairs" rests on a faulty premise

As explained above, WARF's argument about pairs—in conjunction with its analysis of the three "tiers"—mis-reads the intrinsic evidence and mis-applies the law. Perhaps more fundamentally, WARF's argument rests on a faulty premise.

Professor Dally opines:

> **Because the '752 Patent was the first to describe the organization of a dynamic multi-bit prediction associated with a LOAD instruction**, and has clearly described this organization in the specification, **the claim should cover this broader interpretation and not be limited to** the narrower interpretation of a prediction table including a STORE column, and **predictions based on load/store pairs** of the preferred embodiment.

(Dally Decl. ¶ 60.)

The crux of Professor Dally's statement is also the crux of WARF's claim construction argument: the patent would still be valid even if it were not limited to load/store pairs; hence it should receive a "broader interpretation."

Even presuming that the patent would be valid under a broader interpretation (a conclusion with which Intel disagrees), this would not be a basis on which to expand the meaning of its claims. When intrinsic evidence defines the subject matter of an invention to include certain features, it is irrelevant that the invention might have been valid even if defined or claimed without those features. *See Phillips*, 415 F.3d at 1312 (the claims of a patent define the invention); *Atofina v. Great Lakes Chem. Corp.*, 441 F.3d 991, 998 (Fed. Cir. 2006) ("we have held the patentees to the scope of what they ultimately claim, and we have not allowed them to assert that claims should be interpreted as if they had surrendered only what they had to" during prosecution of the patent) (internal quotation omitted); *cf. SciMed Life Sys., Inc. v. Advanced Cardiovascular Sys., Inc.*, 242 F.3d 1337, 1341 (Fed. Cir. 2001) ("Where the

specification makes clear that the invention does not include a particular feature, the feature is deemed to be outside the reach of the claims . . . .").

In sum, WARF's argument on pairs comes to this: the named inventors could have written the patent to be broader than they did. This argument, even if accurate (which Intel disputes), is irrelevant. Here, the claims and intrinsic evidence establish that the "invention" is limited to a data speculation circuit that identifies mis-speculations caused by load/store pairs.

3. "mis-speculation"

'752 Claim Language	WARF's Proposed Construction	Intel's Proposed Construction
1. In a processor capable of executing program instructions in an execution order differing from their program order, the processor further having a data speculation circuit for detecting data dependence between instructions and detecting a **mis-speculation** where a data consuming instruction dependent for its data on a data producing instruction of earlier program order, is in fact executed before the data producing instruction, a data speculation decision circuit comprising:	**mis-speculation:** where a LOAD instruction, that is dependent for its data on a STORE instruction appearing earlier in program order, is in fact executed before the STORE instruction.	**mis-speculation:** where an instruction has been in fact executed prematurely and erroneously

WARF did not address this issue in its Opening Brief. Intel thus continues to rely on its explanations provided in its Opening Brief, at pages 35-36.

4. "predictor"

'752 Claim Language	WARF's Proposed Construction	Intel's Proposed Construction
(a) a **predictor** receiving a mis-speculation indication from the data speculation circuit to produce a prediction associated with the particular data consuming instruction and based on the mis-speculation	**predictor:** a circuit that receives a mis-speculation indication from the data speculation circuit to produce a prediction	**predictor:** a circuit that receives a mis-speculation indication from the data speculation circuit to produce a prediction **based on historical mis-speculation indications**

indication; and		

The parties' disagreement is whether to include the final phrase "based on historical mis-speculation indications" in the claim construction of "predictor."

Intel's proposed construction is correct because the specification and file history define a predictor that bases its prediction on historical mis-speculations. *See* 4:1-4 ("predictor based on **the past history of mis-speculations**"); Tab 11 to Intel's Opening Markman Br., First Amendment, p. 2 ("the predictor" determines data dependency "by examining previous instances of mis-speculation.").

WARF does not address "predictor" in its Opening Brief. However, in its discussion of "prediction," WARF posits that "predictions . . . are based on *several* historical mis-speculations." (WARF Br. at 32, emphasis original.) Given WARF's agreement to the underlying feature—that claim 1 claims a prediction system based on "several historical mis-speculations"—the language proposed by Intel should be included in the construction of "predictor."[2]

[2] Alternatively, the extra language could be included in the construction of "prediction," for example, a "prediction" is "a value *based on historical mis-speculations* indicating the likelihood . . . ." Regardless, this language should be included in the definition of either "predictor" or "prediction."

5. "prediction"

'752 Claim Language	WARF's Proposed Construction	Intel's Proposed Construction
(a) a predictor receiving a mis-speculation indication from the data speculation circuit to produce a **prediction** associated with the particular data consuming instruction and based on the mis-speculation indication; and	**prediction:** a *dynamic multi-bit* value which indicates the likelihood that the data speculative execution of a LOAD instruction will result in a mis-speculation	**prediction:** a value indicating the likelihood that data speculative execution of a load/store *pair* will result in a mis-speculation

The parties' disagreement with respect to "prediction" centers on two issues: (1) load/store "pairs" and (2) "dynamic multi-bit."

a. load/store pairs

This issue is addressed above. Only two points bear additional discussion.

First, WARF criticizes Intel's proposed construction of "prediction" on the grounds that "it does not make sense to refer to speculative execution of a load/store pair," because speculative execution applies only to load instructions. (WARF Br. at 25.) WARF's argument is based on semantics, and does not address any substantive concern. The '752 patent specification refers to a load/store **pair** as the thing that causes mis-speculation: "mis-speculations can be attributed to a few static LOAD/STORE pairs" (3:51-5); "one load/store pair causes a data mis-speculation" (3:55); "the prediction table is checked to see whether the load/store pair causing the mis-speculation is in the prediction table already" (12:65-67). *See* Clark Decl., ¶ 37. The load/store pair causes a mis-speculation when the load in the pair has executed before the store in the pair. *See* Clark Decl., ¶ 37. Similarly, "speculative execution of a load/store pair" means that the load in the pair has executed before the store in the pair. *See* Clark Decl., ¶ 37. The patent specification thus contradicts WARF's argument that it "does not make sense" to adopt this phraseology. *See* Clark Decl., ¶ 37.

Second, as noted in Intel's Opening Brief at 41-42, the patent specification defines the prediction as being based on dependency between "two instructions," a reference to the load/store pair. *See* 14:1-6. *See* Clark Decl., ¶ 38. WARF's proposed construction and the argument in its Opening Brief overlook this lexicography.

b. dynamic and multi-bit

WARF's argument with respect to "dynamic" and "multi-bit" confirms that the words should not be added to the claim construction. WARF is trying to pack these words with meaning and functionality that is not present in the words "dynamic" or "multi-bit." If these terms are inserted into the meaning of "prediction," nothing will be accomplished because an additional Markman hearing will be required to define "dynamic" and "multi-bit."

In support of its proposed construction of "dynamic," WARF cites Dally Decl. ¶ 56, which posits that the prediction is dynamic because it is "updated to reflect the history of this particular load instruction." (WARF Br. at 30-31.) In other words, "dynamic" means "updated," according to WARF. But, the word "dynamic" does not inherently mean "updated," so the word does not itself convey the meaning that WARF (allegedly) intends. Moreover, the "updated" meaning is already captured elsewhere in the claim, and in language more accurate than "updated." For example, the parties agree that the term "prediction associated with the particular data consuming instruction . . . " means "based on historical mis-speculations." The term "dynamic" as a synonym for "updated" thus adds no meaning, and creates the possibility for inaccuracy and confusion.

For "multi-bit," WARF says that the '752 "invention" does not "constitute an on/off decision," and that "[o]nly a multi-bit value can capture a range of predictions." WARF cites Dally Decl. ¶¶ 45 and 56 in support. (WARF Br. at 31.) Dally Decl. ¶ 45 contrasts the alleged invention to prior art in which a "single mis-speculation will prevent the LOAD from being

reordered ahead of the STORE indefinitely," and Dally Decl. ¶ 56 says that the prediction "indicates degrees of likeliness rather than all-or nothing likeliness." *See* Clark Decl., ¶ 39. *See also* WARF Br. at 32 ("predictions . . . are based on several historical mis-speculations").

If WARF's point is that speculative execution of a load instruction is not stopped after one mis-speculation (i.e., **multiple** mis-speculations are required before speculation is prevented), then Intel agrees. *See* Clark Decl., ¶ 39. But this meaning is not captured by the word "multi-bit." *See* Clark Decl., ¶ 39.

Further, the meaning is captured elsewhere in the claim in language that is more clear and accurate than "multi-bit": (a) in "data speculation circuit" and "in fact executed" which emphasize that the system is based on actual mis-speculations ("in fact"); (b) in "predictor" and "prediction associated with the particular data consuming instruction . . . " which require that the predictor acts on "historical mis-speculations"; and, (c) in the "predetermined range," which means that multiple mis-speculations must push the prediction into a "range" of values before speculation is disabled.

Moreover, contrary to WARF's argument, a single-bit prediction could accomplish the functionality that WARF invokes. *See* Clark Decl., ¶ 40. A bit carries the value 0 or the value 1. *See* Clark Decl., ¶ 40. If the system were set to disable speculation after two mis-speculations, then a single bit could implement the system: the first mis-speculation would cause the load/store pair to be entered into the prediction table with a prediction value of "0," and the second mis-speculation would cause the load/store pair to receive a "1" value in the table. *See* Clark Decl., ¶ 40. The value "1" in the table would indicate that there had been two mis-speculations and that the system had hit its threshold. *See* Clark Decl., ¶ 40. As just described, this 0/1 prediction could be implemented using a single bit. *See* Clark Decl., ¶ 40. Indeed, the

'752 patent contemplates this outcome by noting that the entry in the prediction table starts at zero. (11:32, "normally the prediction starts at zero.") *See* Clark Decl., ¶ 40.

It also bears repeating (*see* Intel Opening Br. at 42) that the patent specification never describes the prediction as being "multi-bit." *See* Clark Decl., ¶ 41. This is a term invented by WARF for the litigation. WARF's expert Professor Dally argues that, "the '752 Patent was the first to describe the organization of a dynamic, multi-bit prediction. . . ." (Dally Decl. ¶ 60.) But this is incorrect: the patent specification does not "describe" a "multi-bit prediction," and indeed does not use the term "multi-bit" at all. Moreover, nothing in the patent specification or the claims is written to the level of detail that would concern how many bits are used for the prediction. *See* Clark Decl., ¶ 41.

III. CONCLUSION

For the foregoing reasons, Intel respectfully submits that the Court adopt Intel's proposed claim constructions for the '752 patent.

/s/ Richard Bolton
Richard Bolton
BOARDMAN LAWFIRM LLP
State Bar No. 1012552
One South Pinckney Street, Fourth Floor
Madison, WI 53703
Tel: (608) 283-1789
Fax: (608) 283-1709
rbolton@boardmanlawfirm.com

Of Counsel:

William F. Lee (admitted *pro hac vice*)
Donald R. Steinberg (admitted *pro hac vice)*
Stephen M. Muller (admitted *pro hac vice*)
Victor F. Souto (admitted *pro hac vice*)
WILMER CUTLER PICKERING
HALE AND DORR LLP
60 State Street
Boston, Massachusetts 02109
Tel: (617) 526-6000
Fax:(617) 526-5000

Attorneys for Defendant and Counterclaimant
INTEL CORPORATION

Dated: July 24, 2008

CERTIFICATE OF SERVICE

I hereby certify that on July 24, 2008, copies of the foregoing document were served on the following attorneys at the addresses and in the manner indicated:

BY EMAIL AND FEDERAL EXPRESS:

Michelle Marie Umberger
Lissa Koop
Heller Ehrman LLP
One East Main Street
Ste. 201
Madison, WI 53703-5118

Robert T. Haslam
Anupam Sharma
Heller Ehrman LLP
275 Middlefield Road
Menlo Park, CA 94025

/s/ Richard Bolton